(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－105500

(43)公開日　平成10年(1998)４月24日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	13/00	３５７	Ｇ０６Ｆ	13/00	３５７Ｚ
Ｈ０４Ｌ	12/40		Ｈ０４Ｌ	11/00	３２１
	12/28			11/20	Ｃ

審査請求　未請求　請求項の数８　ＯＬ　（全９頁）

(21)出願番号　特願平9－243266

(22)出願日　平成９年(1997)９月９日

(31)優先権主張番号　０８／７１１１８９

(32)優先日　1996年９月９日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　596092698
ルーセント　テクノロジーズ　インコーポレーテッド
アメリカ合衆国．07974－0636　ニュージャーシィ，マレイ　ヒル，マウンテン　アヴェニュー　600

(72)発明者　ヴラディミール　ネプスティル
アメリカ合衆国　80303　コロラド，ボールダー，クアラ　ドライヴ　4905

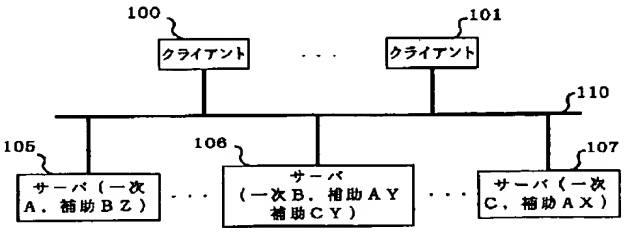
(74)代理人　弁理士　岡部　正夫　(外11名)

(54)【発明の名称】　ネットワーク・サーバの動的再構成

(57)【要約】

【課題】　本発明は、インターネットやイントラネットのアーキテクチャ等のクライアントーサーバシステムのアーキテクチャに関し、特に、サーバの処理負荷を分散する技術を提供する。

【解決手段】　本発明のクライアントーサーバシステムは、クライアント要求を処理する複数のサーバからなり、あるサーバ上の処理負荷が過剰でない間は、全クライアント要求を処理し、該少なくとも１つの第１のサーバ上の処理負荷が過剰になった場合には、ある特定のクライアント要求のみを処理して他のクライアント要求は他のサーバに対して自動的にリダイレクトし、そしてリダイレクト先のサーバはそのリダイレクションに自動的に応答して、リダイレクトされたクライアント要求を処理することを特徴とする。



【特許請求の範囲】

【請求項１】　クライアント・サーバ・システムであって、

クライアント要求を処理する複数のサーバからなり、

該複数のサーバのうちの少なくとも１つの第１のサーバは、第１の情報及び該第１の情報に関連する第２の情報を有し、該少なくとも１つのサーバ上の処理負荷が過剰でない間は、該第１の情報を必要とする該クライアント要求の部分と該第２の情報を必要とする該クライアント要求の部分とを処理し、該少なくとも１つの第１のサーバ上の処理負荷が過剰になったことに応答して、第２の情報を必要とする該クライアント要求の部分を処理することなく第１の情報を必要とする該クライアント要求の部分を処理し、処理をする少なくとも１つの第２のサーバに対して第２の情報を必要とする該クライアント要求の部分を自動的にリダイレクトするものであり、そして、

該複数のサーバのうちの該少なくとも１つの第２のサーバは、該第１および第２の情報のうちの第２の情報を有しており、該リダイレクションに自動的に応答して、該第２の情報を必要とする該クライアント要求の該リダイレクトされた部分を処理することを特徴とするクライアント・サーバ・システム。

【請求項２】　請求項１に記載のシステムにおいて、該少なくとも１つの第２のサーバは、該第１の情報を有することなく該第２の情報を有しており、該第１の情報を必要とするクライアント要求の部分を処理することなく、第２の情報を必要とする該クライアント要求のリダイレクトされた部分を処理することを特徴とするシステム。

【請求項３】　請求項１に記載のシステムにおいて、

該少なくとも１つの第１のサーバ上の処理負荷が過剰でない間は、該第１の情報は該少なくとも１つの第１のサーバの該第２の情報を示す第１のページまたはオブジェクトを含んでおり、該少なくとも１つの第１のサーバ上の処理負荷が過剰になったことに応答して、該少なくとも１つの第２のサーバの該第２の情報へ示す第２のページまたはオブジェクトを含むことを特徴とするシステム。

【請求項４】　請求項１に記載のシステムにおいて、

該少なくとも１つの第１のサーバ上の処理負荷が所定の限界値を超えていない間は、該少なくとも１つの第１のサーバは該クライアント要求の両方の部分を処理し、該少なくとも１つの第１のサーバ上の処理負荷が所定の限界値を超えたことに応答して、該第２の情報を必要とする該クライアント要求の部分を処理することなく該第１の情報を必要とする該クライアント要求の部分を処理し、該第２の情報を必要とする該クライアント要求の部分を自動的にリダイレクトすることを特徴とするシステム。

【請求項５】　請求項４に記載のシステムにおいて、

該所定の限界値は第１の所定の第１の限界値であり、該少なくとも１つの第１のサーバ上の処理負荷が所定の第２の限界値以下になったことに応答して、該少なくとも１つの第１のサーバは、該第２の情報を必要とする該クライアント要求の部分をリダイレクトすることを自動的に止め、該クライアント要求の両方の部分の処理を回復することを特徴とするシステム。

【請求項６】　請求項１に記載のシステムにおいて、

該少なくとも１つの第１のサーバはさらに、該第１の情報に関連付けられている第３の情報をさらに有し、少なくとも１つのサーバ上の処理負荷が所定の第１の限界値を超えていない間は、該第１の情報を必要とする該クライアント要求の部分と、該第２の情報を必要とする該クライアント要求の部分と、該第３の情報を必要とする該クライアント要求の部分とを処理し、該少なくとも１つの第１のサーバ上の処理負荷が該所定の第１の限界値を初めて超過したことに応答して、該第２の情報を必要とする該クライアント要求の部分を処理することなく該第３の情報を必要とする該クライアント要求の部分と第１の情報を必要とする該クライアント要求の部分とを処理し、該第２の情報を必要とする該クライアント要求の部分を、処理をする該少なくとも１つの第２のサーバへ自動的にリダイレクトし、そして該少なくとも１つの第１のサーバ上の処理負荷が該所定の第１の限界値を二度目に超えたことに応答して、該第２の情報を必要とする該クライアント要求の部分および該第３の情報を必要とする該クライアント要求の部分を処理することなく該第１の情報を必要とする該クライアント要求の部分を処理し、該第３の情報を必要とする該クライアント要求の部分を該複数のサーバの処理する少なくとも１つの第３のサーバへ自動的にリダイレクトし、

該複数のサーバのうちの該少なくとも１つの第３のサーバは該第１、第２、および第３の情報のうちの第３の情報を有し、該少なくとも１つの第１のサーバ上の処理負荷が該所定の第１の限界値を二度目に超えた時に、該第３の情報を必要とする該クライアント要求の該リダイレクトされた部分を処理し、そして、

該少なくとも１つの第１のサーバ上の処理負荷が所定の第２の限界値以下に最初に落ちたことに応答して、該少なくとも１つの第１のサーバは更にリダイレクションを自動的に止めて、該第２及び第３の情報を必要とする該クライアント要求の部分の処理を回復し、そして該少なくとも１つの第１のサーバ上の処理負荷が所定の第２の限界値以下に二度目に落ちたことに応答して、リダイレクションを自動的に止めて、該第２および第３の情報のうちの１つを必要とする該クライアント要求の部分の処理を回復することを特徴とするシステム。

【請求項７】　クライアント要求を処理する複数のサーバからなるクライアントーサーバ　システムを操作する

方法であって、該方法が、
該複数のサーバのうちの少なくとも１つの第１のサーバにおいて、該少なくとも１つの第１のサーバ上の処理負荷が過剰でない間は、該第１の情報を必要とする該クライアント要求の部分と該第１の情報に関連付けられている第２の情報を必要とする該クライアント要求の部分の両方を処理する段階からなり、該少なくとも１つの第１のサーバは該第１の情報と該第２の情報の双方を有しており、該システムは更に、
該少なくとも１つの第１のサーバ上の処理負荷が過剰になってきたことに応答して、該少なくとも１つの第１のサーバの中の第２の情報を必要とする該クライアント要求の部分を処理することなく該第１の情報を必要とする該クライアント要求の部分を処理する段階と、
該少なくとも１つの第１のサーバ上の処理負荷が過剰になってきたことに応答して、
該第２の情報を必要とする該クライアント要求の部分を複数のサーバのうちの該少なくとも１つの第１のサーバへ自動的にリダイレクトする段階とからなり、該少なくとも１つの第２のサーバは第２の情報を有しており、該方法は更に、
該リダイレクションに応答して、該少なくとも１つの第２のサーバにおいて該第２の情報を必要とする該クライアント要求のリダイレクトされた部分を自動的に処理する段階からなることを特徴とする方法。
【請求項８】　請求項７に記載の方法においてさらに、該少なくとも１つの第１のサーバ上の処理負荷が所定のレベル以下に落ちたことに応答して、該第２の情報を必要とする該クライアント要求の該部分をリダイレクトすることを自動的に止める段階と、
該少なくとも１つの第１のサーバ上の処理負荷が該所定のレベル以下に落ちたことに応答して、該少なくとも１つのサーバの中の該クライアント要求の該部分の両方の処理を自動的に回復する段階とからなることを特徴とする方法。
【発明の詳細な説明】
【０００１】
【発明の分野】本発明は、インターネットやイントラネットのアーキテクチャなどの、情報ネットワーク・アーキテクチャに関する。
【０００２】
【発明の背景】インターネットなどの情報ネットワークにおいて、クライアントと呼ばれるユーザのコンピュータが、サーバと呼ばれる情報プロバイダのコンピュータから情報を要求し、そのサーバは要求された情報をそのクライアントに対して供給する。インターネット上での情報の格納、検索、および転送のための事実上の標準（de-facto standard）であるワールド・ワイド・ウエブ（World Wide Web：WWW）において、情報はページの形式で供給される。ページはテキスト、グラフィック、スクリプト、または他の形式で表現される情報のディスプレイの全画面である。ページは１つまたはそれ以上のオブジェクトを含む。オブジェクトはユニホーム・リソース・ロケータ（Uniform Resource Locator：ＵＲＬ）と呼ばれる、それ自身のネットワーク・アドレス（ユニークな単独のアドレスであることが好ましい）を有する情報要素である。例えば、ページはディスプレイ画面上で１つまたはそれ以上のテキスト・オブジェクト、１つまたはそれ以上の画像オブジェクト、およびフレーム・オブジェクトによって定義されるレイアウトの中に表示される１つまたはそれ以上のスクリプト・オブジェクトを含むことができる。

【０００３】通常、サーバは自分が提供する情報およびサービスに対する入り口（entry point）として役立つメイン・ページを備えている。このページは通常、同じサーバによってサービスされる他のページおよびオブジェクト（例えば、グラフィック画像、ビデオ／オーディオ／テキスト・ファイル）などを示している。
【０００４】一般に、クライアントがサーバにアクセスする時、サービスはそのクライアントに対してメイン・ページを提供し、次にクライアントと対話してそのクライアントに必要な追加の情報またはサービスを提供する。サーバにアクセスするクライアントの数が増加するにつれて、サーバの処理負荷が増加し、その性能は実質的に劣化する。従って、ユーザがサーバに対して要求を出してからその要求がサーバによって満足されるまでの間の遅延時間が増加することになる。
【０００５】サーバの過負荷を避けるために、一般に管理者はそのサーバを手動で再構成しなければならず、その対象のサーバ上の負荷を軽くするために、要求のいくつかを他のサーバに対してリダイレクトする。いくつかのサービス・プロバイダはサービスされる情報の複製を複数のサーバの中に記憶し、異なるサーバが異なる要求に対して、例えば、総当たり（round-robin）を基礎としてサービスする。それによって複数のサーバに要求の負荷を分散する。これにはいくつかの欠点がある。まず第１に、管理者の手動での介入は遅くて不十分であり、誤りを起こしやすく、そして即座には行われないことが多い。第２に、総当たりを基礎として要求にサービスするために複数のサーバを使うことによって、要求が比較的少数である期間においてサーバの利用率が低下することになり、従って、それは非効率的である。さらに、この方式ではすべてのサーバの情報が各サーバ上に複製されている必要があり、サーバは共通のデータに対する共通のキャッシュを利用することができない。
【０００６】
【発明の概要】本発明は、従来の技術のこれらの問題点および他の問題点を解決することを指向している。一般に、本発明によると、クライアント要求を処理するために、一次サーバが使用する情報の一部分が１つまたはそ

れ以上の補助の待機サーバ上に複製され、クライアントのサービスに対する需要が増加すると、一次サーバ上の処理負荷が過剰になり、一次サーバは情報の複製された部分を必要とするクライアント要求の部分の処理を自動的に補助サーバに肩代わりさせる。サービスに対する需要が減少し、一次サーバが過小負荷状態になると、クライアント要求全体を一次サーバが自動的にサービスするように回復する。

【０００７】本発明の利点は、その時点での処理負荷に基づいて、人間の介入なしに、負荷の分散および負荷の共有が自動的に発生することを含む。サーバ全体の中から１台のサーバまたは１グループのサーバだけが任意の時点において、情報の個々の部分（例えば、ページ、またはオブジェクト、ページまたはオブジェクトのグループ）にサービスし、それによって情報の効率的なキャッシングが可能となる。サービスに対するクライアントの需要が大幅に変動する場合であっても、比較的一様な応答時間がクライアントに対して提供される。さらに、待機サーバが一次サーバのクライアントにサービスしていない間、その処理パワーを他の処理作業のために使うことができる。例えば、サービスに対する需要のピークの時刻が一次サーバのクライアントとは異なっている他のクライアントにサービスし、それによってサーバを効率的に利用することができる。

【０００８】本発明の第１の側面によれば、クライアント・サーバ・システムはクライアント要求を処理するために複数のサーバを含み、複数のサーバのうちの少なくとも１つの第１のサーバが、第１の情報を必要とするクライアント要求の部分および第２の情報を必要とするクライアント要求の部分を処理するために、第１の情報およびその第１の情報に関連付けられた第２の情報を有している。その少なくとも１つのサーバ上の処理負荷が過剰でない間、例えば、所定の第１の限界値を超えていない間、その少なくとも１つの第１のサーバがクライアント要求の両方の部分を処理する。少なくとも１つの第１のサーバ上の処理負荷が過剰になったことに応答して、その少なくとも１つの第１のサーバは第２の情報を必要とするクライアント要求の部分を処理せずに、第１の情報を必要とするクライアント要求の部分を処理し、第２の情報を必要とするクライアント要求の部分を処理するために、少なくとも１つの第２のサーバに対して自動的にリダイレクトする。複数のサーバのうちの少なくとも１つのサーバがその第２の情報を有し、第２の情報を必要とするクライアント要求のリダイレクトされた部分をそのリダイレクションに応答して自動的に処理する。好ましくは、少なくとも１つの第１のサーバ上の処理負荷が、例えば、所定の第２の限界値以下に落ちたことに応答して、その少なくとも１つの第１のサーバはその第２の負荷を必要とするクライアント要求の部分をリダイレクトすることを止め、そのクライアント要求の両方の部分の処理を回復する。

【０００９】本発明の第２の側面によれば、クライアント要求を処理するための複数のサーバを含むクライアントーサーバ・システムを動作させる方法は、次のステップを含む。複数のサーバのうちの少なくとも１つの第１のサーバ上の処理負荷が過剰でない間、その少なくとも１つの第１のサーバは第１の情報を必要とするクライアント要求の部分と、その第１の情報に関連付けられた第２の情報を必要とするクライアント要求の部分の両方を処理し、その少なくとも１つの第１のサーバは第１の情報と第２の情報の両方を持っている。少なくとも１つの第１のサーバ上の処理負荷が過剰になってきたことに応答して、その少なくとも１つのサーバは第２の情報を必要とするクライアント要求の部分を処理せずに、第１の情報を必要とするクライアント要求の部分を処理し、第２の情報を必要とするクライアント要求の部分を、複数のサーバのうちの少なくとも１つの第２のサーバに対して自動的にリダイレクトする。そのリダイレクションに応答して、その少なくとも１つの第２のサーバは第２の情報を必要とするクライアント要求のリダイレクトされた部分を自動的に処理し、その少なくとも１つの第２のサーバは第２の情報を持っている。好適には、少なくとも１つのサーバ上の処理負荷が所定の限界値以下になった時、その少なくとも１つのサーバは第２の情報を必要とするクライアント要求の部分をリダイレクトすることを自動的に止め、クライアント要求の両方の部分の処理を回復する。

【００１０】本発明のこれらおよび他の利点および特徴は図面と一緒に本発明の例示としての実施形態の次の記述から、より明らかになる。

【００１１】

【発明の詳細な記述】図１は例示としての情報ネットワーク（この例においてはインターネット）を示している。それは複数のクライアント１００乃至１０１および複数のサーバ１０５乃至１０７がインターネットのネットワーク構造１１０によって相互に接続されているものを含む。各サーバ１０５はプロセッサおよびメモリを備えているコンピュータであり、その中でプロセッサはそのメモリの中に格納されている制御プログラムを実行して、そのメモリの中に格納されているサービスおよびデータを提供する。各サーバ１０５乃至１０７はそれぞれ、情報Ａ乃至Ｃのデータベースに対する一次サーバである。しかし、本発明によると、データベースに対する一次サーバである以外に、各サーバ１０５乃至１０７は１つまたはそれ以上の他のサーバのデータベースの一部分に対する第２のサーバ、すなわち、サポートしているサーバでもある。図１の例において、サーバ１０５はサーバ１０６のデータベースＢの部分ＢＺに対する補助サーバであり、サーバ１０６はサーバ１０５のデータベースＡの部分ＡＹおよびサーバ１０７のデータベースＣの

一部分ＣＷに対する補助サーバであり、そしてサーバ１０７はサーバ１０５のデータベースＡの一部分ＡＸに対する補助サーバである。

【００１２】図２は本発明を理解するのに重要であるサーバ１０５乃至１０７のメモリ２０５乃至２０７の内容をそれぞれ示している。メモリ２０５はデータベースＡ２１０と、データベースＢ２２０の部分ＢＺ２２１の複製２１１とを格納している。メモリ２０６はデータベースＢ２２０と、データベースＣ２３０の部分ＣＷ２３１の複製２２２と、データベースＡ２１０の部分ＡＹ２１３の複製２２３とを格納している。そして、メモリ２０７はデータベースＣ２３０と、データベースＡ２１０の部分ＡＹ２１２の複製２３２とを格納している。

【００１３】データベースＡ２１０はデータベースＡ２１０の中の他のページＡ２（２４３）乃至Ａ４（２４５）に対するリンクを含む一次メイン・ページＡ１（２４０）を含む。さらに、ページＡ３（２４４）、Ａ４（２４５）はそれぞれ他のページＡ５（２４６）およびＡ６（２４７）乃至Ａ７（２４８）に対するリンクもそれぞれデータベースＡ２１０の中に含む。データベースＡ２１０はデータベース２１０の中のページＡ２（２４３）乃至Ａ３（２４４）に対するリンクを含む第２のメイン・ページＡ１’（２４１）をさらに含むが、メモリ２０５の中の部分ＡＸ２１２のページＡ４（２４５）に対するリンクを含む代わりに、メモリ２０７の中の複製部分ＡＸ２３２の複製ページＡ４（２４５）に対するリンクを含む。さらに、データベースＡ２１０はデータベース２１０の中のページＡ２（２４３）に対するリンクを含む第３のメイン・ページＡ１”（２４２）をさらに含むが、メモリ２０５の中の部分ＡＸ２１２およびＡＹ２１３のページＡ３（２４４）およびＡ４（２４５）に対するリンクを含む代わりに、メモリ２０６および２０７のそれぞれの中の複製部分ＡＹ２２３およびＡＸ２３２の複製ページＡ３（２４４）およびＡ４（２４５）に対するリンクをそれぞれ含む。

【００１４】データベースＢ２２０はオブジェクトＢ２（２５１）に対するリンクおよびデータベースＢ２２０の中の別のページＢ３（２５２）に対するリンクを含んでいるメイン・ページＢ１（２５０）を含む。Ｂ３（２５２）はデータベースＢ２２０の中の他のオブジェクトまたはページＢ４（２５８）乃至Ｂ５（２５９）に対するリンクをさらに含む一次ページである。データベースＢ２２０はメモリ２０６の中のデータベースＢ２２０の部分ＢＺ２２１の中のページＢ４（２５８）乃至Ｂ５（２５９）に対するリンクを含む代わりに、メモリ２０５の中の複製部分ＢＺ２２１の複製オブジェクトおよび／またはページＢ４（２５８）乃至Ｂ５（２５９）に対するリンクを含む第２のページＢ３’（２５３）をさらに含む。

【００１５】データベースＣ２３０はデータベースＣ２３０の部分ＣＷ２３１の中のデータ・オブジェクトＣ２６２に対するリンクを含む、メイン・ページの一次フォーマット・オブジェクトＣ１（２６０）を含む。データベースＣ２３０はメモリ２０７の中のデータベースＣ２３０の部分ＣＷ２３１の中のデータ・オブジェクトＣ２（２６２）に対するリンクを含む代わりに、メモリ２０６の中の複製部分ＣＷ２２２の複製オブジェクトＣ２（２６２）に対するリンクを含む、メイン・ページの二次的なフォーマット・オブジェクトＣ１’（２６１）をさらに含む。

【００１６】サーバ１０５乃至１０７は処理負荷が最も重くなる時点がそれぞれ異なると仮定される。例えば、サーバ１０５は夕方において最も頻繁に使われる可能性があり、サーバ１０６は週日において最も頻繁に使われる可能性があり、サーバ１０７は週末に使われることが最も多い可能性がある。最初に、すべてのサーバ１０５乃至１０７は普通に動作する。すなわち、サーバ１０５はデータベースＡ２１０からの情報に対するすべての要求にサービスし、サーバ１０６は最初にデータベースＢ２２０からの情報に対するすべての要求にサービスし、そしてサーバ１０７はデータベースＣ２３０からの情報に対するすべての要求にサービスする。それは従来の方法で、一次ページおよびオブジェクト２４０、２５２、２６０、またはこれらのいずれか一方を使って行われる。また普通のように、各サーバ１０５乃至１０７は、例えば、単位時間当たりにサービスされたアクセス（要求）の回数の形で、現在の処理負荷の記録を維持する。

【００１７】本発明を理解するために重要なサーバ１０５乃至１０７の動作が、それぞれ図３乃至図５の中に示されている。この従来の動作の他に、各サーバは所定の負荷限界値で初期化される負荷制御プログラムを実行する。図３に示されているように、サーバ１０５は現在の負荷が高負荷上限値「Ａ１’ｈｉｇｈ」を超えているかどうかをステップ３００において繰返しチェックする。上限値を超えていなかった場合、それはサーバ１０５が過負荷になっていないことを意味し、従って、サーバ１０５はステップ３００にとどまっている。負荷が上限値を超えていた場合、それはサーバ１０５が過負荷であることを意味し、従って、サーバ１０５は一次ページＡ１（２４０）に対して二次ページＡ１’（２４１）をステップ３０２において置き換える。これはデータベースＡ２１０の部分ＡＸ２１２からの情報に対するそれ以降のすべての要求がサーバ１０７に対して向けられるようにする効果がある。サーバ１０７はこれらの要求を複製部分ＡＸ２３２から従来の方法でサービスする。従って、サーバ１０５が過負荷になると、その処理負荷のいくつかはサーバ１０７によって肩代わりされる。

【００１８】ステップ３０２に続いて、サーバ１０５は高負荷の下限値「Ａ１’ｌｏｗ」がその現在の処理負荷を超えているかどうかをステップ３０４においてチェッ

クする。超えていた場合、それはサーバ１０５が過小負荷状態にあることを意味し、従って、サーバ１０５はステップ３０６において、二次ページＡ１’（２４１）に対して一次ページＡ１（２４０）を置き換える。これは初期動作を回復する効果を持つ。その場合、サーバ１０５はデータベースＡ２１０からの情報に対するすべての要求にサービスしている。次に、サーバ１０５はステップ３００へ戻る。

【００１９】限界値Ａ１’ｌｏｗが現在の負荷を超えていないことがステップ３０４において判定された場合、サーバ１０５は現在の負荷が再びＡ１’ｈｉｇｈの限界値を超えていないかどうかをステップ３０８においてチェックする。超えていなかった場合、それはサーバ１０５が過負荷状態でないことを意味し、従って、サーバ１０５はステップ３０４へ戻る。Ａ１’ｈｉｇｈの限界値を超えていた場合、それはサーバ１０５が再び負荷になっていることを意味し、従って、サーバ１０５はステップ３１０において、二次ページＡ１’（２４１）に対して三次ページＡ１’（２４２）を置き換える。これはデータベースＡの部分ＡＹ２１３からの情報に対するそれ以降のすべての要求がサーバ１０６に対して向けられるようにする追加の効果がある。サーバ１０６は複製部分ＡＹ２２３から従来の方法でこれらの要求にサービスし、それによってサーバ１０５によって行われるはずであった処理負荷のいくつかを肩代わりし、従って、サーバ１０５の負荷が軽減される。

【００２０】ステップ３１０に続いて、サーバ１０５はＡ１’ｌｏｗの限界値がサーバ１０５の現在の処理負荷を超えているかどうかを、ステップ３１２においてチェックする。超えていなかった場合、サーバ１０５はステップ３１２の状態にとどまる。超えていた場合、それはサーバ１０５が過小負荷になっていることを意味し、従って、サーバ１０５は三次ページＡ１’（２４２）に対して二次ページＡ１’（２４１）をステップ３１４において置き換える。これはステップ３１０においてサーバ１０６に対して転送された処理負荷の部分をサーバ１０５が戻す効果がある。次に、サーバ１０５はステップ３０４へ戻る。

【００２１】サーバ１０６と１０７の動作は似ている。図４に示されているように、サーバ１０６は現在の処理負荷が高負荷の上限値「Ｂ３’ｈｉｇｈ」を超えているかどうかを、ステップ４００において繰返しチェックする。超えていなかった場合、サーバ１０６は過負荷状態ではなく、ステップ４００にとどまる。超えていた場合、サーバ１０６は過負荷状態であり、従って、サーバ１０６は一次ページＢ３（２５２）に対して二次ページＢ３’（２５３）をステップ４０２において置き換える。これはデータベースＢ２２０の部分ＢＺ２２１からの情報に対するそれ以降のすべての要求がサーバ１０５に対して向けられるようにする効果がある。サーバ１０５はこれらの要求に対して複製部分ＢＺ２１１から従来の方法でサービスする。それによってサーバ１０６の負荷が解放される。

【００２２】ステップ４０２に続いて、サーバ１０６は高負荷の下限値「Ｂ３’ｌｏｗ」がその現在の処理負荷を超えているかどうかをステップ４０４においてチェックする。超えていなかった場合、サーバ１０６はステップ４０４にとどまる。超えていた場合、それはサーバ１０６が過小負荷状態であることを意味し、従って、サーバ１０６は二次ページＢ３’（２５３）に対して一次ページＢ３（２５２）をステップ４０６において置き換える。これは初期の動作を回復する効果があり、サーバ１０６はデータベースＢ２２０からのすべての要求にサービスしている。次にサーバ１０６はステップ４００へ戻る。

【００２３】図５に示されているように、サーバ１０７はその現在の処理負荷が負荷限界値「Ｃ１’」を超えているかどうかを、ステップ５００において繰返しチェックする。超えていなかった場合、サーバ１０７は過負荷状態ではなく、ステップ５００にとどまる。超えていた場合、サーバ１０７は過負荷状態であり、従って、それは一次オブジェクトＣ１　２６０に対して二次オブジェクトＣ１’（２６１）をステップ５０２において置き換え、それによってその処理負荷の一部をサーバ１０６へ転送する。ステップ５０２に続いて、サーバ１０７は負荷限界値「Ｃ１’」がその現在の処理負荷を超えているかどうかをステップ５０４において繰返しチェックする。超えていなかった場合、サーバ１０７はステップ５０４にとどまる。超えていた場合、それはサーバ１０７が過負荷状態にはなっていないことを意味し、従って、サーバ１０７はステップ５０６において、二次オブジェクトＣ１’（２６１）に対して一次オブジェクトＣ１（２６０）を置き換える。これによってその初期動作を回復する。次にサーバ１０７はステップ５００へ戻る。

【００２４】もちろん、上記の例示としての実施形態に対する各種の変更および修正が可能であることは、この分野の技術に熟達した人にとっては明らかである。例えば、一次および二次のページまたはオブジェクトの両方を格納する代わりに、一次ページまたはオブジェクトが「高速（on-the-fly）」で（すなわち、リアルタイムで）二次ページまたはオブジェクトに変換されるようにすること、あるいはその逆が行われるようにすることができる。同様に、データベースの複製部分が補助サーバ上にあらかじめ格納されている代わりに、データベースの部分が「高速」で補助サーバに対して複製されて分配されるようにすることができる。さらに、単位時間当たりのアクセスの回数以外の測定値および限界値を使って、一次サーバとの間の負荷の肩代わりまたは戻しの処理を行うかどうかを判定することができる。これらの測定値および限界値は過去において同様な時刻において経

験された負荷に基づいて、将来の負荷を推定する予測アルゴリズムなどの前方監視（forward-looking）とすることができる。さらに、メイン・サーバは待機サーバから現在の処理負荷データを要求することができ、そしてこれらのデータをそれらの待機サーバに対する肩代わりの処理を行うかどうかの判定に組み込むことができる。そのような変更および修正は本発明の精神および範囲から逸脱することなしに、そしてその利点を減らすことなしに行うことができる。従って、そのような変更および修正は特許請求の範囲によってカバーされることが意図されている。

【図面の簡単な説明】

【図１】本発明の例示としての実施形態を含む情報ネットワークのブロック図である。

【図２】図１の情報ネットワークのサーバのメモリの部分的内容のブロック図である。

【図３】図１の情報ネットワークのサーバの異なるものの部分的動作のフローチャートである。

【図４】図１の情報ネットワークのサーバの異なるものの部分的動作のフローチャートである。

【図５】図１の情報ネットワークのサーバの異なるものの部分的動作のフローチャートである。

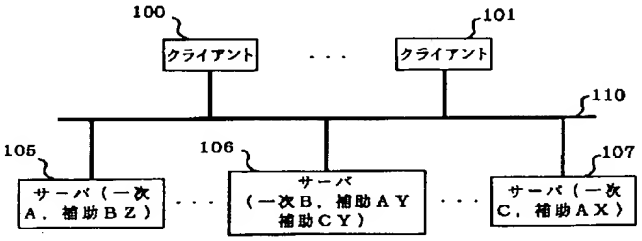
【符号の説明】

２０５　メモリ

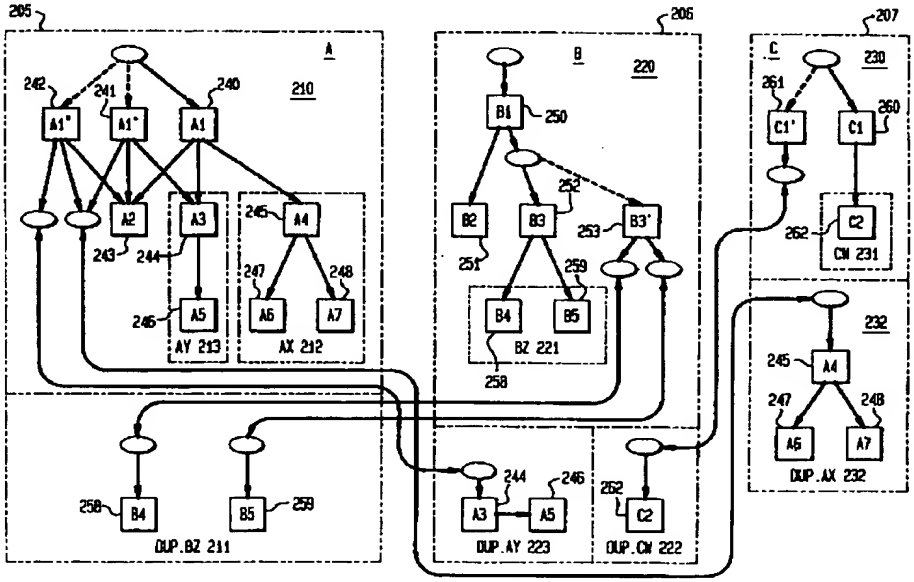
２０６　メモリ

２０７　メモリ

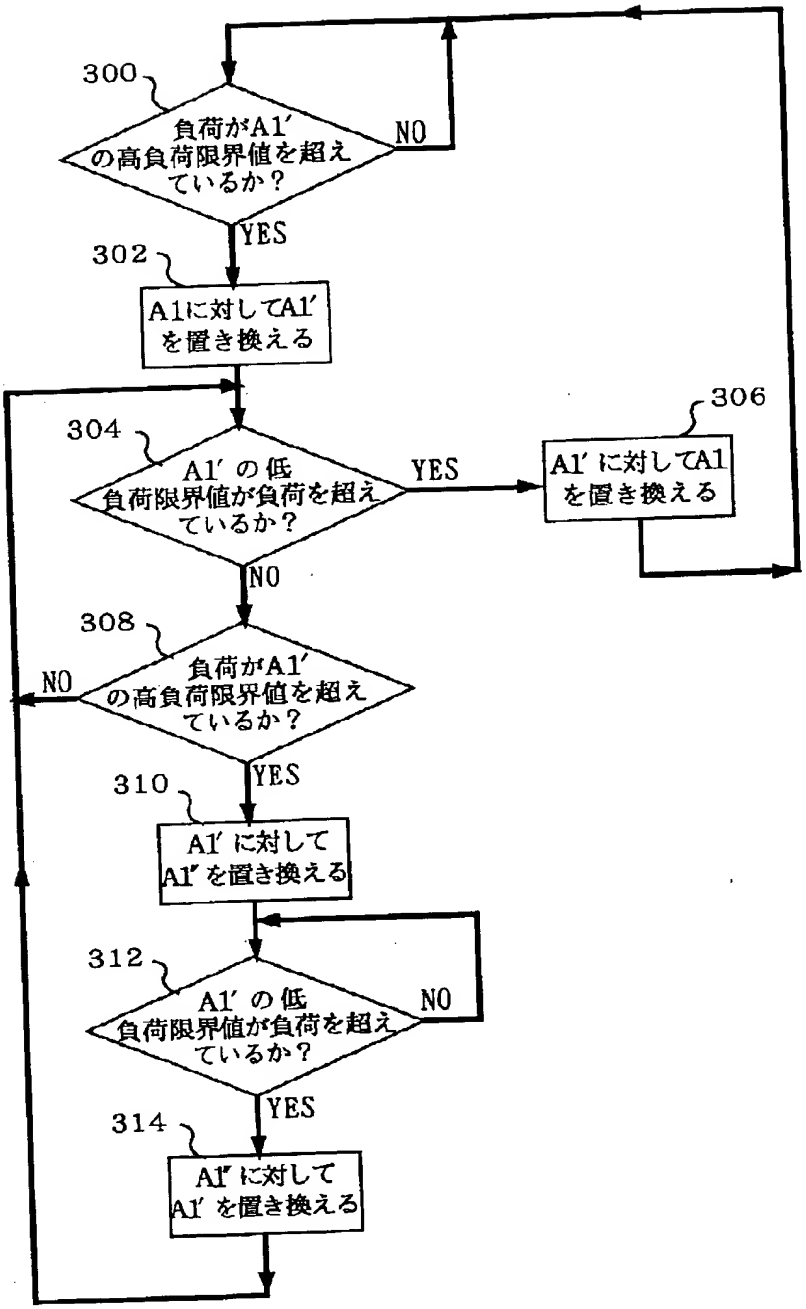
【図１】



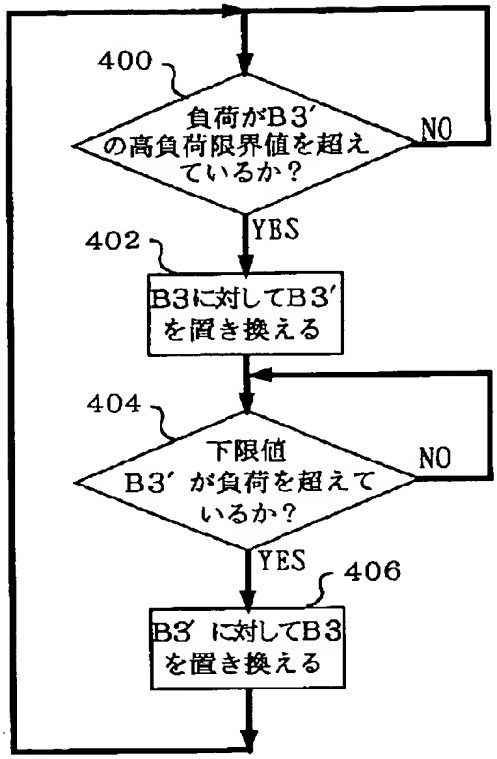
【図２】



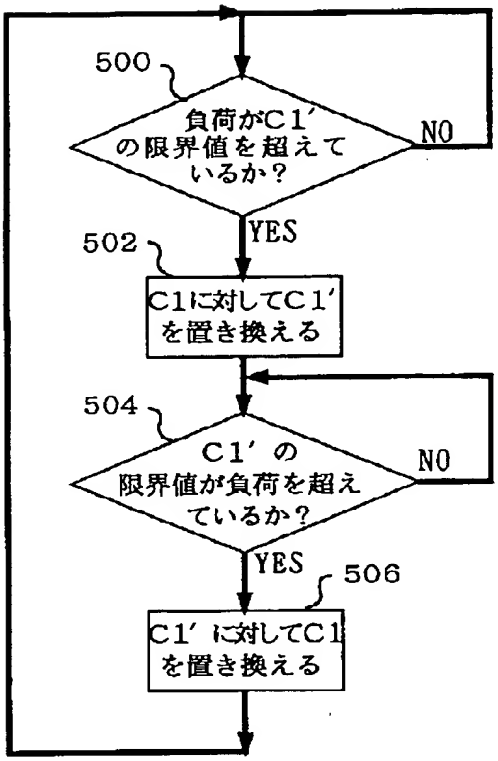
【図３】


300
負荷がA1′の高負荷限界値を超えているか？
NO
YES
302
A1に対してA1′を置き換える
304
A1′の低負荷限界値が負荷を超えているか？
YES
306
A1′に対してA1を置き換える
NO
308
負荷がA1′の高負荷限界値を超えているか？
NO
YES
310
A1′に対してA1′を置き換える
312
A1′の低負荷限界値が負荷を超えているか？
NO
YES
314
A1′に対してA1′を置き換える

【図４】


400
負荷がB3′の高負荷限界値を超えているか？
NO
YES
402
B3に対してB3′を置き換える
404
下限値B3′が負荷を超えているか？
NO
YES
406
B3′に対してB3を置き換える

【図５】


500
負荷がC1′の限界値を超えているか？
NO
YES
502
C1に対してC1′を置き換える
504
C1′の限界値が負荷を超えているか？
NO
YES
506
C1′に対してC1を置き換える